на правах рукописи
УДК 517

Волков Леонид Михайлович

МОДЕЛИ И АЛГОРИТМЫ ОБРАБОТКИ ИНФОРМАЦИИ В ПРОГРАММНЫХ КОМПЛЕКСАХ ЭЛЕКТРОННОГО ДОКУМЕНТООБОРОТА

05.13.18 — математическое моделирование, численные методы и комплексы программ

Автореферат
диссертации на соискание ученой степени кандидата физико-математических наук

Екатеринбург — 2006

Работа выполнена на кафедре алгебры и дискретной математики Уральского государственного университета им. А.М.Горького.

Научный руководитель:	доктор физико-математических наук, профессор В.А. Баранский.
Официальные оппоненты:	доктор технических наук, профессор А.А. Захаров, кандидат физико-математических наук, доцент В.Б. Костоусов.
Ведущая организация:	Нижегородский государственный университет им Н.И. Лобачевского.

Защита состоится _____ 2006 года в ____ часов на заседании диссертационного совета К 212.286.01 по присуждению ученой степени кандидата физико-математических наук при Уральском государственном университете им. А.М. Горького по адресу: 620083, Екатеринбург, пр. Ленина 51, комн. 248.

С диссертацией можно ознакомиться в научной библиотеке Уральского государственного университета им. А.М. Горького.

Автореферат разослан «___» _____ 2006 года.

Ученый секретарь диссертационного совета,
доктор физико-математических наук,
профессор

В.Г. Пименов

Общая характеристика работы

Актуальность представляемой диссертации. Программный комплекс электронного документооборота — это автоматизированная информационная система, предназначенная для реализации процесса удаленного обмена большими массивами форматированной информации. В наше время, в связи с бурным развитием Интернет-технологий, программные комплексы электронного документооборота находят широчайшее применение во многих сферах человеческой деятельности, и в первую очередь – в процессе электронного взаимодействия государственных структур и хозяйствующих экономических субъектов.

Организация такого взаимодействия является одной из важнейших задач и приоритетов современного информационного общества. Постоянно растут объемы информации, обрабатываемой в информационных системах органов исполнительной власти для повышения качества и эффективности управления государством, и, соответственно, увеличиваются объемы документооборота между бизнес-структурами и государственными органами, уполномоченными законодательством на прием и обработку различного рода данных.

Основной проблемой при этом остается состояние среды, в которой происходит это взаимодействие. Если на обоих концах канала передачи информации в подавляющем большинстве случаев находятся современные автоматизированные информационные системы, умеющие эффективно и качественно обрабатывать получаемую информацию, то сам канал представляет собой бухгалтера предприятия, перемещающегося на общественном транспорте с толстыми папками отчетов, или неспешную почтовую посылку, содержащую, опять же, многочисленные бумаги. Столь явное несоответствие между качеством и пропускной способностью используемых каналов передачи информации и систем обработки данных приводит к тому, что последние, по принципу лимитирующего фактора, оказываются загруженными отнюдь не на полную мощность.

Ресурс существенного повышения качества систем обработки информации, заключается, таким образом, в переводе всего процесса взаимодействия между ними исключительно на электронные рельсы. Именно передача данных в электронном виде по телекоммуникационным каналам связи является единственным естественным способом взаимодействия для современных информационных систем.

Поэтому неудивительно, что уже в течение достаточно длительного времени как в России, так и за рубежом (где эти процессы начались несколькими годами ранее) ставятся и решаются задачи, относящиеся к молодой предметной области под названием «электронное правительство» (eGovernment).

Внедрение систем электронного обмена информацией в масштабах государства требует теоретического обоснования правильности принципов, на которые опираются проекты и технические задания разрабатываемых систем. Проблема заключается в том, что, в связи с «молодостью» всей предметной области, общие принципы проектирования таких программных комплексов не нашли пока систематического понимания и изложения.

На практике создаются и внедряются программные комплексы, ориентированные на решение только части задач электронного документооборота, и не претендующие, таким образом, на гибкость и масштабируемость. Например, в представленной на отечественном рынке системе «Комита-Отчет» (разработчик – ЗАО «Комита», Санкт-Петербург) решается только задача физической передачи форматированных данных по каналам связи, но остается открытой проблема подтверждения валидности данных. А в популярной системе «Такском-Спринтер» (ООО «Такском», Москва) отсутствуют механизмы для работы со всеми историческими состояниями форм электронных документов (см. [5]). Такие проблемы характерны не только для России: в 2005 году Германии из-за ошибок в проектировании провалился национальный проект по представлению налоговой отчетности через Интернет, так как программное обеспечение не было рассчитано на масштабирование в условиях лавинообразного увеличения нагрузки.

Актуальность исследования принципов проектирования программных комплексов защищенного и юридически значимого электронного документооборота обусловлена, таким образом, тем, что

- благодаря применению таких систем становится возможным повышение эффективности государственного управления, за счет ускорения поступления информации в автоматизированные информационные системы государственных органов;
- реальное внедрение систем будет происходить в условия бурного роста количества абонентов и развития возможностей самих программных комплексов, поэтому необходимо заранее прогнозировать вектор развития информационной модели и дать теоретические оценки пределам возможностей этих систем;
- одновременно требуется внедрение целого ряда систем документооборота (например, в России потребителями информации выступают налоговая служба, фонды социального и медицинского страхования, служба государственной статистики, таможенная служба, Пенсионный фонд, служба по тарифам, служба финансового мониторинга, служба экологического мониторинга, региональные финансово-бюджетные службы – все со своими требованиями к процедурам обработки информации), и, следовательно, имеется потребность в обобщении принципов их проектирования, и в построении универсальной модели обработки данных, на базе которой была бы возможна интеграция этих систем;
- объемы документооборота (в России – до 2.5 миллиардов документов в год) и требования по доступности исторических версий документов (срок хранения по отдельным видам документов – до 75 лет) представляют собой вызов к производительности и масштабируемости информационных систем, равного которому история разработки программного обеспечения еще не знала.

В диссертационной работе автором дается теоретическое обоснование ряду аспектов архитектуры программных комплексов электронного документооборота, а именно

- предлагаются новые, эффективные алгоритмы обработки информации в таких комплексах, связанные с обеспечением возможности работы с

информационными массивами больших объемов, хранящими все исторические состояния данных;

- предлагается и обосновывается компонентная модель обработки и хранения информации в узле системы, обеспечивающая автоматическую проверку целостности данных;
- описываются универсальные, масштабируемые модели информационных потоков и преобразований, обеспечивающие гибкость системы, заключающуюся в настройке новых форм и видов документов на расширяемом языке метаописаний, без привлечения программистов.

При этом для решения задачи, имеющей большое практическое значение, удается привлечь и развить известные научные результаты теоретической информатики. Так, в своем исследовании структур данных, обеспечивающих доступ ко всем своим историческим состояниям, мы опираемся на работы Р.Е.Тарджана и соавторов ([7]), а говоря о принципах построения объектной модели обработки информационных потоков, мы развиваем методологию проектирования Г.Буча и Э. Гаммы ([6], [3]). Организация хранения хронологических структур данных в реляционных СУБД требует привлечения теории последних, изложенной, например, в монографии К.Дж.Дейта ([4]).

Цель работы: исследуя процессы обработки, хранения и контроля целостности данных, циркулирующих в среде автоматизированного программного комплекса электронного документооборота

- установить общие закономерности, которым подчиняются данные процессы;
- создать на основе этих закономерностей математическую модель информационных потоков в программных комплексах электронного документооборота;
- разработать алгоритмы работы с данными, обеспечивающие эффективное хранение и обработку информации в узлах системы документооборота;

- предложить методологию применения данной математической модели при проектировании, разработке и внедрении прикладных программных комплексов.

Методика исследования опирается на результаты дискретной математики, теории объектно-ориентированного проектирования и реляционных баз данных. Для получения теоретических результатов осуществляются дедуктивные рассуждения, а построение практических моделей производится методами математического моделирования.

Научная новизна. Все основные результаты диссертационной работы являются новыми. В частности, это алгоритмы построения и проверки целостности хронологических деревьев, модель информационных потоков в программном комплексе электронного документооборота и программная компонентная модель обработки данных KDOM.

Практическая значимость. Основные результаты диссертационной работы имеют как теоретическое, так и практическое значение. Результаты, полученные в области алгоритмов обработки хронологических деревьев, могут быть использованы при исследовании широких классов хронологических структур данных. Построенная модель обработки информации в программном комплексе электронного документооборота применима при проектировании широкого класса используемых на практике программных комплексов.

Под руководством и с участием диссертанта в компании «СКБ Контур» была разработана основанная на изложенных в диссертации моделях и алгоритмах система защищенного электронного документооборота «Контур-Экстерн». Основное назначение системы – организация передачи налоговой и бухгалтерской отчетности от налогоплательщиков и налоговых органов. Практическое внедрение системы «Контур-Экстерн» состоялось, к 1 августа 2006 года, в 80 регионах Российской Федерации (из 88). Количество абонентов системы на 1 августа 2006 года превысило 125000. Ежеквартальные объемы документооборота в системе превышают 1.5 миллиона электронных документов. В течение всего времени промышленной эксплуатации системы (с января 2003 года) удвоение количества абонентов происходило не более чем за год.

Апробация работы. Основные результаты диссертации докладывались на конференциях «Инфофорум. Безопасность информации в современном обществе» (Москва, 2003, 2004, 2005), «Проблемы региональной информатизации» (Ханты-Мансийск, 2002, 2003, 2004, 2005), «Информатизация налоговых органов» (Московская область, 2004, 2005, 2006), «Инфоком. Решения для электронной России» (Москва, 2003), на Уральской конференции молодых ученых (Екатеринбург, 2003), российско-германском семинаре BitKom (Ганновер, 2005), российско-корейском семинаре по информационным технологиям (Екатеринбург, 2005), на научных семинарах кафедры алгебры и дискретной математики УрГУ. Основные результаты диссертации были также опубликованы в виде тезисов в сборниках трудов вышеперечисленных конференций ([15-18]).

Система «Контур-Экстерн», спроектированная автором и разработанная под его руководством, отмечена премией конкурса «Лучший инновационный проект Уральского федерального округа», проводившегося в 2002 году под руководством Полномочного представителя Президента Росссийской Федерации в УрФО П.М.Латышева. Система «Контур-Экстерн» является также лауреатом национальной конференции в области информационной безопасности «Инфофорум-2003» (в номинации «Технология года») и «Инфофорум-2005» (в номинации «Проект года»).

Публикации. Основные результаты диссертации опубликованы в работах [8]-[11], [17], [18], список которых приводится в конце автореферата.

Авторские права на созданные диссертантом программные комплексы оформлены в виде свидетельств [13], [14]. Разработанная модель обмена информацией в системе защищенного юридически значимого электронного документооборота защищена Патентом Российской Федерации [12].

В патенте и авторском свидетельстве, оформленных совместно с Э.Р. Шифманом, диссертанту принадлежат принципиальная модель обмена информацией и алгоритмы обработки данных, Э.Р. Шифману – постановка задачи и определение целевой функциональности системы.

Структура и объем работы. Диссертация состоит из введения, двух глав, заключения и списка литературы, содержащего 71 наименование. Материал диссертации изложен на 163 страницах, снабжен 12 иллюстрациями.

Автор работы выражает благодарности

научному руководителю, профессору кафедры алгебры и дискретной математики Уральского государственного университета, доктору физико-математических наук Виталию Анатольевичу Баранскому за многочисленные ценные наблюдения, замечания и обсуждения в процессе подготовки диссертации;

техническому директору компании «СКБ Контур» Эдуарду Романовичу Шифману за вдохновляющие обсуждения, помогшие пробросить мостик от теории к практике.

Содержание работы

Во введении формулируется задача моделирования информационных потоков в программном комплексе электронного документооборота, и приводятся аргументы в пользу актуальности этой задачи. В качестве объекта исследования выбирается

Система электронного документооборота — программный комплекс, предназначенный для передачи информации по телекоммуникационным каналам связи между территориально удаленными информационными массивами.

Актуальность исследования программных комплексов электронного документооборота вытекает из возможности их применения в автоматизированных системах государственных масштабов, при отсутствии на рынке готовых решений, пригодных по своим потребительским качествам (масштабируемость, настраиваемость, отказоустойчивость) для внедрения в промышленную эксплуатацию при объемах документооборота порядка миллиона документов в месяц и выше.

Те или иные аспекты архитектуры систем электронного документооборота неоднократно были предметом рассмотрения в научной и технической литературе.

В литературном обзоре особое внимание уделено отечественным и зарубежным функциональным прообразам систем электронного документооборота – системам управления документами (таким, как «Евфрат» - [1]) и системам передачи данных по каналам связи (см., например, [5]). Анализ результатов предшественников позволяет выделить подзадачи, решение которых позволило бы на практике создать и внедрить программный комплекс электронного документооборота, достаточно производительный и гибкий для эксплуатации в составе государственной системы электронного документооборота с хозяйствующими субъектами. Среди этих подзадач особое внимание привлекают те, которые оставались нерешенными на момент начала автором его исследовательской работы. В частности, выясняется, что задача построения систем электронного документооборота требует решения двух проблем, ранее не исследовавшихся применительно к практике разработки программных комплексов электронного документооборота:

- разработка алгоритмов и информационных моделей для организации хранения и обработки больших массивов данных, изменяющихся во времени;
- разработка математической модели автоматизированной обработки информационных потоков, представленных в различных форматах и формах, устойчивой по отношению к изменению форматов.

Две главы работы посвящены рассмотрению вышеуказанных открытых проблем построения систем электронного документооборота.

В первой главе диссертации рассматриваются задачи, связанные с организацией хранения и обработки больших массивов данных, изменяемых во времени. Ключевым определением главы является

Определение 1.1. Типом данных «хронологический(ое, ая) T» называется отображение, которое каждому моменту времени из некоторого интервала ставит в соответствие структуру данных типа T, при этом для различных моментов времени носители структур, которые порождаются отображением, совпадают.

Дальнейший анализ концентрируется на свойствах хронологических деревьев. С точки зрения практической постановки задачи, оказывается наиболее важным найти оптимальное по скорости выполнения операций навигации

представление хронологических деревьев как структур в памяти, и простой алгоритм восстановления хронологических деревьев как структур в памяти из таблиц реляционных баз данных.

Автором предложена эффективная реализация хронологического дерева как совокупности трех отображений

```
ChTree := {    OID ⇒ (Start ⇒ PID),
               OID ⇒ (Start ⇒ LCID),
               OID ⇒ (Start ⇒ RBID)    };
```

Здесь отображение (Start ⇒ PID) хранит в себе *историю* изменений указателя на родителя данного объекта и представляет собой упорядоченный по ключу Start набор значений идентификатора родителя PID для моментов времени Start. LCID и RBID суть идентификаторы левого ребенка и правого брата данного элемента дерева соответственно (терминология и обозначения заимствованы из [1]).

Наиболее сложной задачей является инициализация рассматриваемой структуры за один проход из списка связей (OID, PID, Start). Для решения этой задачи автором разработан алгоритм, обоснование корректности которого представляет собой один из основных результатов главы. На вход алгоритма поступает тройка (OID, PID, Start); используя указанные в ней данные, необходимо обновить шесть массивов (историй): ParentID, LCID и RBID для объекта PID и то же самое для объекта OID. Здесь трудность заключается, в первую очередь, в пересчете массива указателей на правого брата объекта OID при инициализации хронологического дерева. Для решения этой ключевой подзадачи разработан и обоснован

Алгоритм 1.1.

Шаг 1. Положим S=Start.

Шаг 2. Если S>=t_1, то выход. Иначе, для момента времени S, вычислить левого ребенка объекта PID, пусть это будет объект C_1. Пусть также связь, определяющая, что C_1 — это левый ребенок PID в момент времени S, действует до момента времени s_1. Положим i=1.

Шаг 3. Если C_i = NULL, то перейти к следующему шагу. Иначе, вычислим C_{i+1} как правого брата объекта C_i в момент времени S. Вычислим также s_i как момент времени, до которого действует связь, определяющая, что C_{i+1} — это правый брат объекта C_i в момент времени S.

Шаг 4. Теперь у нас построен список C_1, ... C_k детей объекта PID на момент времени S. Вычислим T = min{s_1, ..., s_k}. Тогда, по построению, T>S и T — это момент времени, до которого список детей объекта PID, построенный на момент времени S, остается неизменным (возможно, что T=∞)

Шаг 5. Найдем такой номер ребенка m, что C_m < OID < C_{m+1}. Перестроим истории изменений правого брата для объектов C_m и OID так, чтобы с момента времени S по момент времени T правым братом объекта C_m был бы OID (это надо сделать, если m>0), а правым братом объекта OID был бы C_{m+1} (возможно, при этом, что C_{m+1}=NULL).

Шаг 6. Положим S=T. Перейдем ко второму шагу.

В силу прикладной необходимости, для хронологических деревьев автором также ставятся и решаются задачи обеспечения целостности. Статическая задача целостности характеризуется следующим вопросом: «Пусть АТД T моделируется структурой данных в памяти D. Дано некоторое состояние структуры D. Верно ли, что это состояние описывает некоторый объект типа T?». Динамическая задача проверки целостности отвечает на другой вопрос: «Определенное состояние данных D описывает объект класса T. В данные внесены некоторые изменения. Верно ли, что новое состояние структуры данных D’ по-прежнему описывает некоторый объект класса T?»

Очевидно, что если найден алгоритм, решающий статическую задачу, то с помощью этого же алгоритма можно решать и динамическую задачу — достаточно лишь применить найденный алгоритм к состоянию данных D’. Поэтому, теоретический и практический интерес представляет такой вопрос:

Пусть состояние D’ в некотором смысле мало отличается от состояния D. Существует ли алгоритм решения динамической задачи, который будет существенно эффективнее статической проверки целостности состояния D’?

Этот вопрос конструктивно решается применительно к хронологическим деревьям. Для этого вводятся следующие понятия:

Определение 1.2. Пусть P —массив Parent некоторого дерева (то есть его элементы проиндексированы натуральным числами от 1 до N и содержат целые значения от 0 до N). *Изменение* в массиве P — это пара целых чисел (OID, PID) таких, что $1 \le OID \le N$ и $0 \le PID \le N$. *Результат применения* изменения (OID, PID) к массиву P — это массив P', заданный соотношениями

$P'[i] = P[i]$, если $i \neq OID$,

$P[i] = PID$, если $i = OID$.

Определение 1.3. Пусть P — массив Parent, задающий ациклический граф. Изменение (OID, PID) будем называть *допустимым*, если результат применения этого изменения P' — также ациклический граф. В противном случае, такое изменение будем называть *недопустимым*.

Определение 1.4. Пусть P — некоторый массив Parent. *Набор изменений* в массиве P — это множество пар целых чисел (OID_i, PID_i), таких, что

- Для каждого i, пара (OID_i, PID_i) — изменение в смысл определения 1.2;
- Если $i \neq j$, то $OID_i \neq OID_j$.

Основным результатом первой главы диссертационной работы является

Теорема 1.1. *Если каждое изменение в наборе недопустимо, то и весь набор недопустим.*

Следствие 1.1. *Если набор изменений допустим, то эти изменения можно применить по одному друг за другом в таком порядке, чтобы после каждого шага граф оставался ацикличным.*

Благодаря этому теоретическому результату, удается построить легко реализуемый на практике динамический алгоритм 1.4 проверки целостности хронологического дерева, существенно более производительный, нежели статический алгоритм в тех случаях, когда различия между последовательными срезами невелики или носят локальный характер. Точную оценку производительности этого алгоритма дает

Теорема 1.2. *Алгоритм 1.4 обеспечивает проверку допустимости набора изменений Changes в массиве Parent и работает за время $O(K^2 \cdot H)$, где H — высота леса, заданного массивом Parent, а K — количество изменений в наборе.*

Таким образом, для проверки целостности «следующего» среза хронологического дерева, который получается из «текущего» среза путем применения к нему набора изменений, имеется два различных алгоритма – статический и динамический. Время работы этих двух алгоритмов зависит от разных параметров: статический алгоритм работает за время $O(N)$, где N — количество элементов в дереве, а динамический алгоритм — за время $O(K^2 \cdot H)$. Вычислительная система, управляющая хронологическим деревом, в состоянии достаточно быстро рассчитывать параметры N, K и H. Таким образом, эта система (менеджер транзакций) в состоянии сама выбирать оптимальный алгоритм для проверки целостности очередного среза хронологического дерева. Такую возможность «интеллектуального» поведения системы удается применить на практике во многих случаях, например при решении задачи проверки целостности всего хронологического дерева.

Во второй главе диссертации рассматриваются вопросы автоматизированной обработки информации, представленной в различных формах и форматах у разных пользователей программного комплекса электронного документооборота. Показано, что наиболее важная подзадача здесь заключается в полной автоматизации преобразования данных из одной формы в другую, которое должно происходить без доработки программного обеспечения комплекса. Основой для рассмотрений является интуитивно понятный тезис о том, что в любой автоматизированной системе электронного документооборота одна и та же информация представляется как в человекочитаемой форме, так и в машиночитаемой. Здесь под человекочитаемой формой представления информации подразумеваются данные, которым сопутствует определенный «фон», позволяющий человеку определить семантику каждого из атомарных значений, а под машиночитаемой формой – поток атомарных значений, снабженных синтаксическими признаками.

Задача преобразования информации из одной формы в другую должна решаться в каждой системе электронного документооборота и, более того, что эта задача является основной задачей таких систем. Цель применения программных комплексов электронного документооборота заключается в снижении объемов бумажного документооборота (который осуществляется полностью в человекочитаемой форме), за счет передачи части функций по обмену документами автоматизированной системе. Следовательно, для достижения цели в первую очередь надо решить задачу переработки данных в форму, пригодную для автоматизированной обработки и обратно, в человекочитаемую форму. Сама переработка информации из одной формы в другую должна производиться автоматизированно, то есть поддержка каждого конкретного формата не должна быть связана с внесением изменений и дополнений в программный код.

Для решения этой задачи обработки информации автор предлагает и обосновывает разработанную им общую математическую модель информационных потоков в программном комплексе электронного документооборота. Модель состоит из четырех основных компонентов (модулей) – storage (хранилище), loader (модуль загрузки данных), saver (модуль выгрузки данных), checker (модуль автоматизированной проверки).

Необходимость разработки внутреннего формата и выделенного хранилища данных обосновывается следующими соображениями:

- система электронного документооборота представляет различным контрагентам одну и ту же информацию в различных формах;
- внутри системы, таким образом, необходимо поддерживать метаинформацию обо всех формах представления циркулирующей в системе информации;
- части этой метаинформации, относящиеся к различным формам представления информации, обязательно различаются, и обязательно имеют достаточно много общего;
- внешние форматы контролируются контрагентами и обслуживают их текущие нужды и потребности, а потому могут быть с течением

времени подвержены существенным изменениям; внутреннее же хранилище системы должно по возможности сохранять стабильность.

Хранилище данных должно обеспечивать следующую функциональность:

- возможность быстрого обхода всей структуры данных и получения произвольных ее подмножеств;
- возможность быстрого сохранения и восстановления данных;
- поддержку транзакций;
- поддержку интерфейсов загрузки информации и обхода.

Для того чтобы обеспечить заполнение хранилища данными, поступающими в систему от контрагентов, необходим модуль загрузки, который является буфером между контрагентами и интерфейсом загрузки хранилища.

Модуль загрузки:

- принимает данные определенного формата из внешнего источника;
- принимает описание формата из хранилища;
- принимает параметры загрузки;
- осуществляет анализ поступивших данных в соответствии с форматом, и преобразует их в поток значений;
- ограничивает (изменяет) поток значений в соответствии с параметрами загрузки;
- направляет результирующий поток значений в хранилище через интерфейс загрузки.

Модуль выгрузки решает задачу формирования внешнего представления данных по внутреннему представлению, содержащемуся в хранилище. Таким образом, его функция является обратной к функции модуля загрузки. Модуль выгрузки является буфером между интерфейсом обхода хранилища и внешними потребителями данных.

Задача модуля выгрузки заключается в следующем:

- Получить от контрагента задание на выгрузку определенных данных;
- Преобразовать это задание в систему команд интерфейса обхода;
- Произвести обход хранилища и получить поток данных, в соответствии с запросом контрагента, а также метаинформацию,

необходимую для представления запрошенных данных в виде, доступном для обработки этим контрагентом;

- Представить поток данных в виде файла воспринимаемого контрагентом формата (если контрагент — автоматизированная система, рис. 1) или в человекочитаемой форме (если контрагент — человек, рис. 2), используя полученную из хранилища метаинформацию.

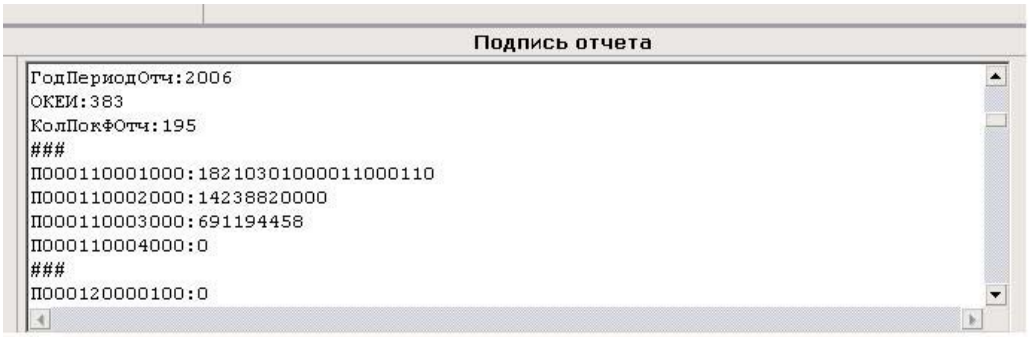
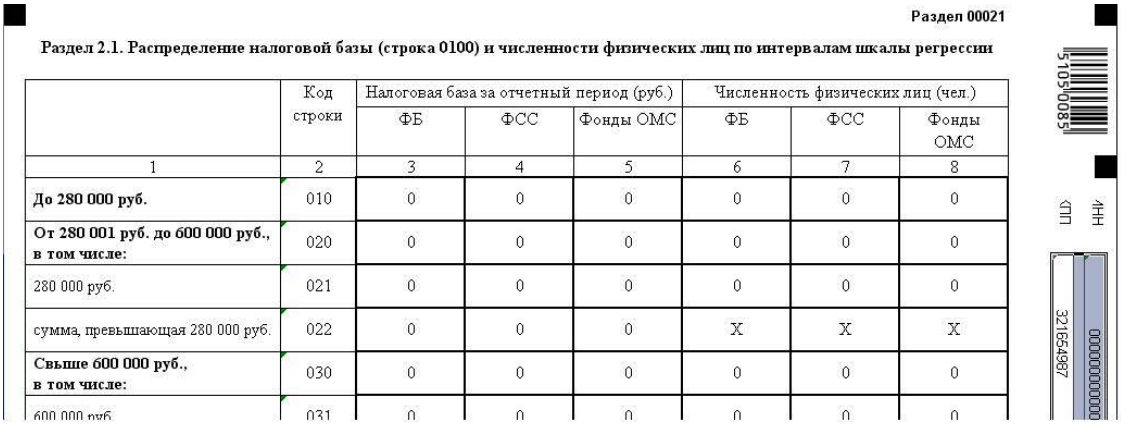


Рис. 1. Результат работы модуля выгрузки – машиночитаемый документ.



Раздел 00021

Раздел 2.1. Распределение налоговой базы (строка 0100) и численности физических лиц по интервалам шкалы регрессии

	Код строки	Налоговая база за отчетный период (руб.)			Численность физических лиц (чел.)		
		ФБ	ФСС	Фонды ОМС	ФБ	ФСС	Фонды ОМС
1	2	3	4	5	6	7	8
До 280 000 руб.	010	0	0	0	0	0	0
От 280 001 руб. до 600 000 руб., в том числе:	020	0	0	0	0	0	0
280 000 руб.	021	0	0	0	0	0	0
сумма, превышающая 280 000 руб.	022	0	0	0	X	X	X
Свыше 600 000 руб., в том числе:	030	0	0	0	0	0	0
600 000 руб.	031	0	0	0	0	0	0

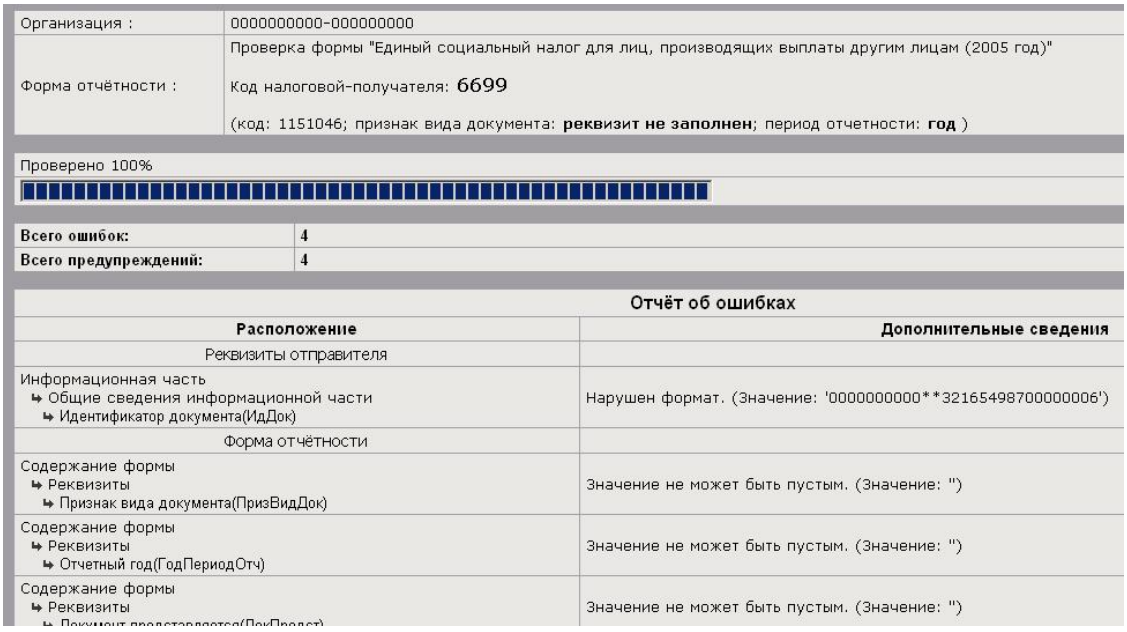
Рис. 2. Результат работы модуля выгрузки – человекочитаемый бланк

Хранилище содержит в себе всю имеющуюся в системе информацию и метаинформацию во внутреннем формате, и может взаимодействовать с внешним миром через модули загрузки и выгрузки информации. В ходе этих процессов содержимое хранилища претерпевает постоянные изменения. Поэтому в системе электронного документооборота должен присутствовать модуль проверки,

отвечающий за поддержание хранилища в соответствии с имеющимися в системе внешними требованиями (ограничениями).

Модуль проверки:

- запускается по факту внесения изменений, с определенной периодичностью или по заказу пользователя системы;
- принимает из хранилища формализованное описание ограничений;
- осуществляет обход хранилища, принимая из него группы реквизитов, затрагиваемые ограничениями, и проверяя выполнение этих ограничений;
- исправляет нарушения ограничений в соответствии с имеющимися в описании ограничений предписаниями;
- формирует протокол своей работы (рис. 3) с описанием обнаруженных нарушений и возвращает его заказчику проверки;



Организация :	0000000000-000000000
Форма отчётности :	Проверка формы "Единый социальный налог для лиц, производящих выплаты другим лицам (2005 год)" Код налоговой-получателя: 6699 (код: 1151046; признак вида документа: **реквизит не заполнен**; период отчетности: **год**)

Проверено 100%

Всего ошибок:	**4**
Всего предупреждений:	**4**

Отчёт об ошибках

Расположение	**Дополнительные сведения**
Реквизиты отправителя	
Информационная часть ↳ Общие сведения информационной части ↳ Идентификатор документа(ИдДок)	Нарушен формат. (Значение: '0000000000**32165498700000006')
Форма отчётности	
Содержание формы ↳ Реквизиты ↳ Признак вида документа(ПризВидДок)	Значение не может быть пустым. (Значение: '')
Содержание формы ↳ Реквизиты ↳ Отчетный год(ГодПериодОтч)	Значение не может быть пустым. (Значение: '')
Содержание формы ↳ Реквизиты ↳ Документ представляется(ДокПредст)	Значение не может быть пустым. (Значение: '')

Рис. 3. Протокол работы модуля проверки в системе «Контур-Экстерн».

В работе показано, что хранилище, модули выгрузки и загрузки являются обязательными компонентами любой правильно спроектированной системы электронного документооборота. Модуль проверки является факультативным, но должен присутствовать в том случае, если требованиями предметной области к

данным выдвигаются определенные ограничения, не описываемые только форматами машиночитаемого представления информации.

На логическом уровне, работа участника системы документооборота с ее модулями обработки и хранения информации ограничивается следующим:

- пользователь имеет возможность передавать данные в модуль загрузки, сообщать параметры загрузки этих данных в хранилище, инициализировать процедуру загрузки данных в хранилище и получать протокол с результатами загрузки;
- пользователь знает о содержимом хранилища в части данных, владельцем которых он является и уверен, что эти данные сохраняются в неизменности между сеансами работы;
- пользователь имеет возможность инициировать проверку состояния хранилища и получать протокол с результатами проверки;
- пользователь имеет возможность инициировать запуск модуля выгрузки, сообщать параметры выгрузки данных и получателя данных.

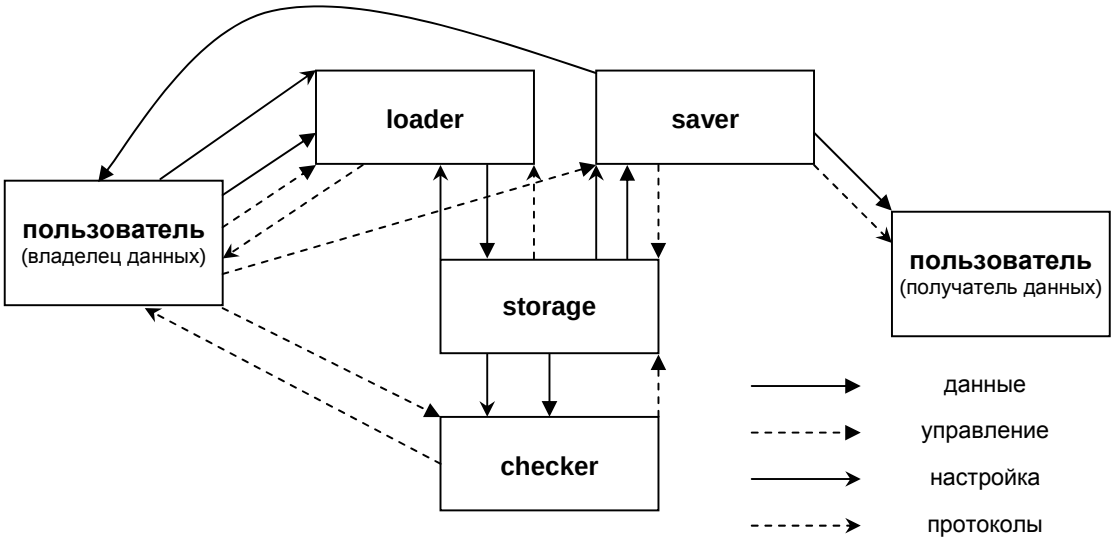


Рис. 4. Соединение модулей обработки информации в системе электронного документооборота

Автором предлагается программная реализация этой математической модели процессов обработки информации в виде компонентной модели KDOM, основанной на XML-метаописаниях. Система KDOM — это набор программных компонентов, выполненных в технологии Microsoft .NET Framework, и взаимосвязей между ними (интерфейсов).

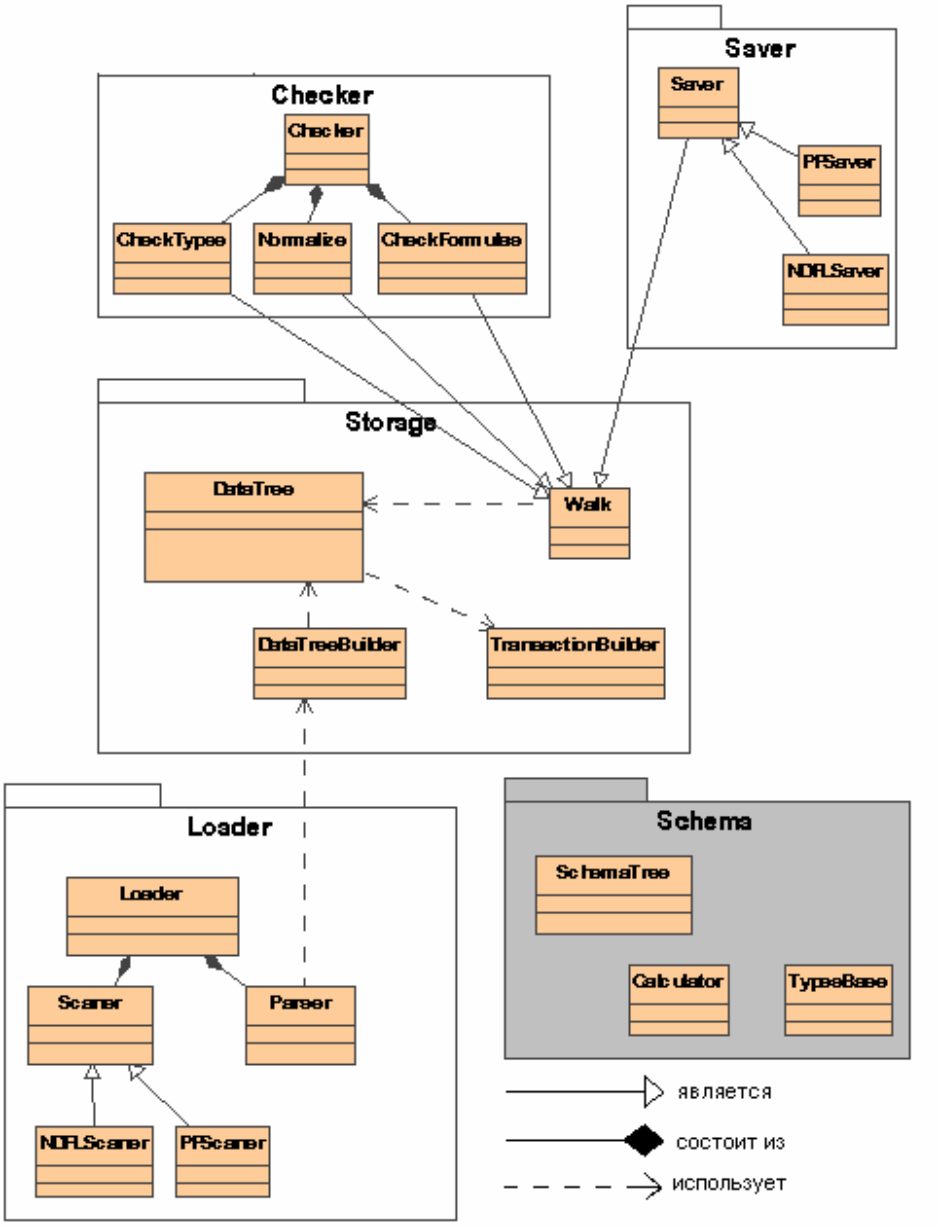


Рис. 5. Схема модели KDOM в нотации UML

Для управления метаинформацией используется набор вспомогательных компонентов, объединенный на рис. 5 логическим блоком Schema. Все компоненты модели KDOM обрабатывают информацию, используя описания форматов, сделанные на специализированном языке. Этот язык построен на базе расширяемого языка разметки XML, который является в настоящее время промышленным стандартом для представления метаданных. Язык метаописаний модели KDOM позволяет описывать синтаксические правила распознавания элементов данных в машиночитаемом файле, синтаксические правила формирования машиночитаемых и человекочитаемых данных в соответствии с форматами внешних потребителей информации, а также описывать исчерпывающие правила форматного и логическо-арифметического контроля валидности информации в хранилище системы.

Важнейшими свойствами системы KDOM являются ее

- расширяемость — разработчик имеет возможность создавать свои реализации компонентов Saver и Scaner для стыковки готовой внутренней архитектуры KDOM с любыми внешними приложениями и пользовательскими интерфейсами;
- гибкость — все описания форматов, проверок, пользовательских типов, дополнительных ограничений целостности, могут быть сделаны в виде XML-схем; механизм управления типами данных позволяет описывать домены произвольной структуры и проводить автоматизированные проверки принадлежности реквизита домену;
- централизованность — все данные и метаданные консолидируются в едином внутреннем хранилище, что обеспечивают быстроту и полноту обработки информации;
- надежность — используются процедуры сохранения, восстановления, контроля целостности и проверки внешних ограничений, обеспечивающие интеллектуальный контроль состояния внутреннего хранилища, а также механизм транзакций, защищающий хранилище от нарушений структуры данных в процессе работы системы.

Таким образом, компонентная модель KDOM является гибко настраиваемой программной средой, которая может быть использована как универсальная объектная модель для разработки любой системы обработки и проверки информации, представляемой в виде формализованных документов.

В заключительном разделе работы содержится резюме о практике внедрения спроектированной автором системы электронного документооборота. Внутренняя архитектура сервера системы «Контур-Экстерн» полностью построена на модели KDOM. Хранилище данных в памяти системы, в процессе обработки информации организуется как хронологический лес, а для длительного хранения упаковывается в СУБД как соответствующая реляционная таблица. Обмен данными между сервером системы и абонентскими терминалами построен с использованием описанного в главе 2 протокола и формата пакета документов.

Таким образом, цели диссертационной работы оказываются достигнутыми: построенные теоретические модели и алгоритмы преобразования и обработки данных в программных комплексах электронного документооборота выдерживают испытание практикой, и позволяют создать прикладную систему, не имеющую аналогов по обслуживаемым объемам документооборота, масштабируемости и гибкости.

Список литературы

1. Арлазаров В.Л., Емельянов Н.Е. Прикладные аспекты построения систем на основе документооборота. // в сб. «Документооборот. Прикладные аспекты». — М.:Институт системного анализа РАН, 2004.
2. Ахо А., Хопкрофт Дж., Ульман Дж. Структуры данных и алгоритмы. — М.: Вильямс, 2003. — 384 с.
3. Гамма Э., Хелм Р., Джонсон Р., Влиссидес Дж. Приемы объектно-ориентированного проектирования. Паттерны проектирования. — СПб.: Питер, 2004. — 368 с.
4. Дейт К.Дж. Введение в системы баз данных. — М.: Вильямс, 2005. — 1328 с.
5. ООО «Такском», Техническая документация на программный комплекс «Спринтер», 2000-2005. — http://www.taxcom.ru/system/technology/.
6. Booch G., Vilot M. The design of the C++ Booch components // Proceedings of the Object-oriented programming systems, languages and applications conference, Ottawa. — ACM Press, 1990, — p. 1-11.
7. Driscoll J.R., Sarnak N., Sleator D.D., Tarjan R.E. Making data structures persistent // J.Comput.System.Sci. — 1989, No. 28, Vol. 1. — p. 86-124.

Работы автора по теме диссертации

8. Волков Л.М. Хронологические структуры данных . Способы представления в памяти. // Екатеринбург: Известия УрГУ. Компьютерные науки. – 2006, №1. – С. 15-25.
9. Волков Л.М. Электронная цифровая подпись в безбумажном документообороте хозяйствующих субъектов и государственных контрольных органов. // М.: PCWeek/RE. - 2002, №43. - С. 20.
10. Волков Л.М. Задачи целостности для хронологических структур данных. // В сб. «Проблемы теоретической и прикладной математики. Труды 34-й региональной молодежной конференции». - Екатеринбург: УрО РАН. - 2003. – С. 250-253.
11. Волков Л.М. Принцип единого окна: как построить систему электронного документооборота между государственными органами и хозяйствующими субъектами. // М.: PCWeek/RE. – 2005, №46. – С. 52-54.
12. Волков Л.М., Шифман Э.Р. Автоматизированная система подготовки и представления отчетности. — Патент на полезную модель №43983, приоритет от 10.02.2005.
13. Волков Л.М., Шифман Э.Р. Система защищенного документооборота «Контур-Экстерн». — Свидетельство о регистрации авторского права на программу для ЭВМ №2004611946 от 23.08.2004.
14. Волков Л.М. Автоматизированное рабочее место приема и обработки информации АРМ «Прием». — Свидетельство о регистрации авторского права на программу для ЭВМ №2004611947 от 23.08.2004.
15. Волков Л.М. Электронный документооборот между предприятиями и государственными контрольными органами. // в сб. «Проблемы региональной информатизации и пути их решения. Сборник трудов научно-практической конференции». - Ханты-Мансийск: Комитет по информационным ресурсам ХМАО. – 2002. – С. 76-79.

16. Волков Л.М. Практика интеграции технологии PKI и прикладных систем электронного документооборота. // «Сборник материалов летней сессии 5-й Всероссийской конференции «Информационная безопасность России в условиях глобального информационного общества» под ред. члена-корреспондента РАЕН А.В.Жукова. – М.: Редакция журнала «Бизнес+Безопасность». – 2003. – С. 105-109.
17. Volkov L. Kontur-Extern — an infractructure for secured digital exchange. // Proceedings of the BitKom Seminar, CeBIT. – Hannover: Deutsche Messe. - 2005. – S. 31-33.
18. Volkov L. Digital data exchange. // Proceedings of the 1st Russian-Korean International Workshop on Mobile and Telecommunication Technology. - Yekaterinburg: Ural State University. - 2005. – P. 70-71.